Evidence Synthesis Infrastructure
Collaborative (ESIC) planning process:
Interim Report

Global SDG Synthesis Coalition (GSDGSC)
Building a Global Evidence Synthesis Community (BGESC)
Pan-African Collective for Evidence (PACE)
Center for Rapid Evidence Synthesis (ACRES)

Working group 3:
Safe & Responsible Use of AI

Stage 4a report:
Deliver: What would make the biggest impact?

Last updated:
23 May 2025

Consultation window:
26-29 May 2025

EXECUTIVE SUMMARY

Working Group 3 (WG3) focuses on the safe and responsible use of AI in evidence synthesis, aligning with ESIC's broader vision to strengthen evidence systems and accelerate progress toward the Sustainable Development Goals (SDGs). The Stage 4a report builds on previous stages: Stage 1 identified existing capabilities and gaps, Stage 2 assessed their maturity, and Stage 3 proposed 50 potential solutions. Stage 4 focuses on making the biggest impact, using a structured prioritization process based on innovation, feasibility, and potential impact.

The Working Group (WG) members initially identified 50 potential solutions aimed at addressing key challenges. To determine which solutions had the highest impact, a structured Delhi survey was conducted, allowing WG members to rate each solution based on feasibility, scalability, and overall effectiveness. The survey used a quantitative rating scale to assess each solution's potential impact, with participants providing detailed rationales for their ratings. Following the analysis of the ratings, responses were aggregated, and an average impact score was calculated for each solution. This allowed for a clear ranking of the most impactful solutions, based on cumulative ratings and expert consensus. **Evidence Synthesis Studio (ESS)** emerged as the highest-rated solution among all WG members, demonstrating a strong agreement on its potential to drive meaningful change. We examined the interconnections among all identified solutions, recognizing patterns and thematic similarities. Solutions with aligned objectives were grouped together to streamline efforts and enhance coherence. Additionally, solutions that required integration to fulfill the overarching vision were strategically clustered, ensuring a unified approach that maximized impact and efficiency. This report presents our **proposed framework** for AI-Digital Evidence Synthesis Tools (AI-DESTs), emphasizing the interconnection between core infrastructure components.

Evidence Synthesis Infrastructure provides a foundational systems, standards, and platforms that support the entire process. CESPIA, the Comprehensive-Evidence Synthesis Plug-in archive for AI-DESTs, serves as both a governance framework and a live inventory for validated AI-DEST tools. It plays an active role in defining validation standards, maintaining datasets, and ensuring interoperability across evidence synthesis platforms. Rather than simply listing validated tools, CESPIA establishes rigorous validation protocols, curates datasets for independent testing, and provides the infrastructure necessary for ongoing assessment of AI-DEST capabilities. The Evidence Synthesis Studio (ESS) serves as a dynamic AI-supported infrastructure, integrating DESTs across various synthesis processes that enables multiple studios to function as adaptable environments for evidence synthesis. ESS allows users to assemble DESTs in a ‘plug and play’ manner, tailoring workflows to their specific review needs. This flexibility ensures that researchers can configure their synthesis processes dynamically, whether working individually in a studio-like setting—similar to R Studio—or within a structured team environment where a predefined set of DEST is curated for broader application by downstream users.

The Evidence Synthesis process and projects encompasses the methodologies, workflows, and best practices that dictate how AI-DEST is applied in research. Best-Practice Guidance, developed by methodologists, provides essential frameworks for trainers and users, ensuring standardization in evidence synthesis. Evaluation and Validation Studie**s** assess the reliability and quality of AI-DEST tools and workflows, maintaining high standards across the ecosystem.

The cohesive ecosystem supports efficient, accurate, and accessible AI-enabled evidence synthesis, grounded in transparency, community input, and technical robustness.

The ESIC planning process is funded by the Wellcome Trust.

Box 1. Glossary of terms

Artificial Intelligence Digital Evidence Synthesis Tools (AI-DESTs): A subset of DESTs that specifically leverage artificial intelligence, machine learning, or natural language processing to automate or enhance evidence synthesis tasks.
Digital Evidence Synthesis Tools (DESTs): Tools that facilitate evidence synthesis processes.
Explainable AI (XAI): processes and methods that allow human users to comprehend and trust the results and outputs created by machine learning algorithms.
Effort: how much resource (time, money, manpower, etc.) as task will require
FAIR data: FAIR data adheres to principles ensuring that data is Findable, Accessible, Interoperable, and Reusable.
GenAI: A class of AI models capable of generating new content such as text, images, or data based on patterns learned from large datasets
Human-in-the-loop (HITL): HITL refers to a process or system design where human oversight, intervention, and collaboration are integrated into Automated Decision-Making Systems at strategic points.
Impact: the magnitude of the benefits a task will bring upon completion
Safe and Responsible Use of AI: Safe and Responsible use of AI in evidence synthesis involves several key principles to ensure ethical, transparent, sustainable, equitable, and effective integration of AI technologies.
Security measures: The management, operational, and technical controls (i.e., safeguards or countermeasures) prescribed for an information system to protect the confidentiality, integrity, and availability of the system and its information.

INTRODUCTION

The Evidence Synthesis Infrastructure Collaborative (ESIC) was established to foster an inclusive and transparent planning framework that supports investment in evidence synthesis infrastructure. This framework will offer an expansive, structured list of potential investments and opportunities for leveraging existing systems.

Working Group 3 (WG3) contributes to this mission by focusing on the Safe and Responsible Use of AI in Evidence Synthesis, aligned with ESIC's broader vision of strengthening evidence systems to address societal challenges and accelerate progress toward the Sustainable Development Goals (SDGs). Using ESIC's double diamond methodology (divergent exploration followed by convergent focus) WG3 frames its work through ethical, technical, and practical lenses to ensure that AI tools serve the evidence ecosystem responsibly and inclusively.

The Stage 4a report builds on our previous work: Stage 1 identified existing capabilities and gaps; Stage 2 assessed the maturity of these capabilities; and Stage 3 proposed 50 potential solutions (Appendix 1). In Stage 4 (Deliver), we address the question: What would make the biggest impact? To answer this, we applied a structured prioritization process based on criteria such as innovation, feasibility, and potential impact, including an impact-effort matrix. Stage 4a refines the original 50 solutions into a focused, high-impact catalog, while Stage 4b will assess implementation costs to support strategic decision-making. Methods for Stage 4 report are presented in Box 2.

Box 2. Methodology

WG3 process: We conducted an anonymous survey among WG3 members using JISC Online Surveys, asking them to rate the 50 proposed solutions from our Stage 3 report on Likert scales (1–10) for each of Impact and Effort, and to provide comments supporting their evaluation. For each proposal we calculated mean scores independently for Impact and Effort, and a combined score. A weakness of this approach is that it assumes ratio-scale properties of the evaluations, so as a sensitivity analysis we complimented this approach using Bayesian ranking; which came to substantially the same conclusions. The results were discussed in WG3 meetings, were through an iterative process, we refined and grouped the solutions.

Impact-Effort matrix: Results were visualized (using R libraries) in an impact-effort matrix, where we used the median and interquartile range (IQR) of the scores to plot the solutions. The IQR was used as a measure of dispersion for each dimension (impact and effort), and the average IQR was inverted (1 / mean IQR) to define the size of each point. Consensus levels were classified as high (IQR ≤ 2), moderate (2 < IQR ≤ 4), and low (IQR > 4).

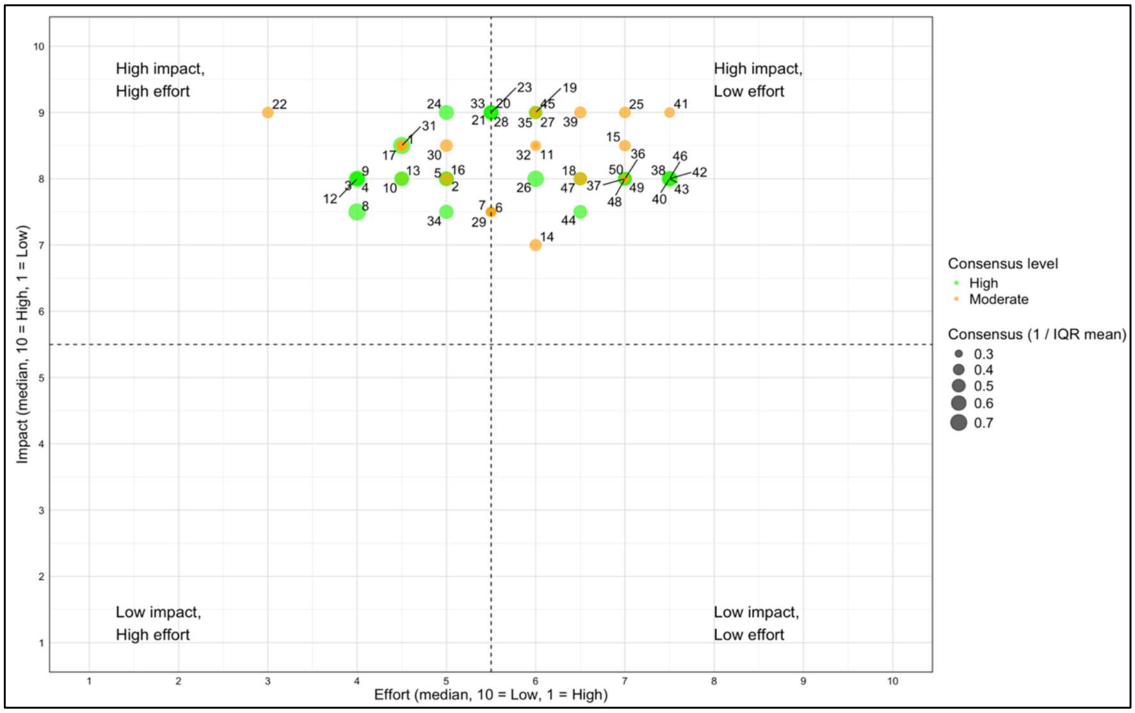
External consultation: We are conducting semi-structured interviews with a range of relevant stakeholders aimed to validate and adjust the proposed strategies. Stakeholders were purposively selected to ensure diversity and representation across gender, region, and sector. Additionally, suggestions received during the open consultation windows of previous reports were considered when identifying interview participants. A dedicated committee was formed to design the interview guide, which was iteratively refined throughout the process. WG3 members conduct the interviews, and key insights are extracted and summarized for analysis.

In this report we present an integrated framework for AI-Digital Evidence Synthesis Tools (AI-DESTs), highlighting the interconnection between core infrastructure components. It presents CESPIA as a live, community-driven validation platform; the ES Data Store as a centralized evidence repository; and the Evidence Synthesis Studio (ESS) as a flexible infrastructure integrating AI across synthesis methods. Complementary components including training platforms, validation studies, and best-practice guidance, build capacity and ensure high standards.

Together, these elements form a cohesive ecosystem that supports efficient, accurate, and accessible AI-enabled evidence synthesis, grounded in transparency, community input, and technical robustness.

IMPACT EFFORT MATRIX TEMPLATE

Figure 1. Impact-Effort Matrix



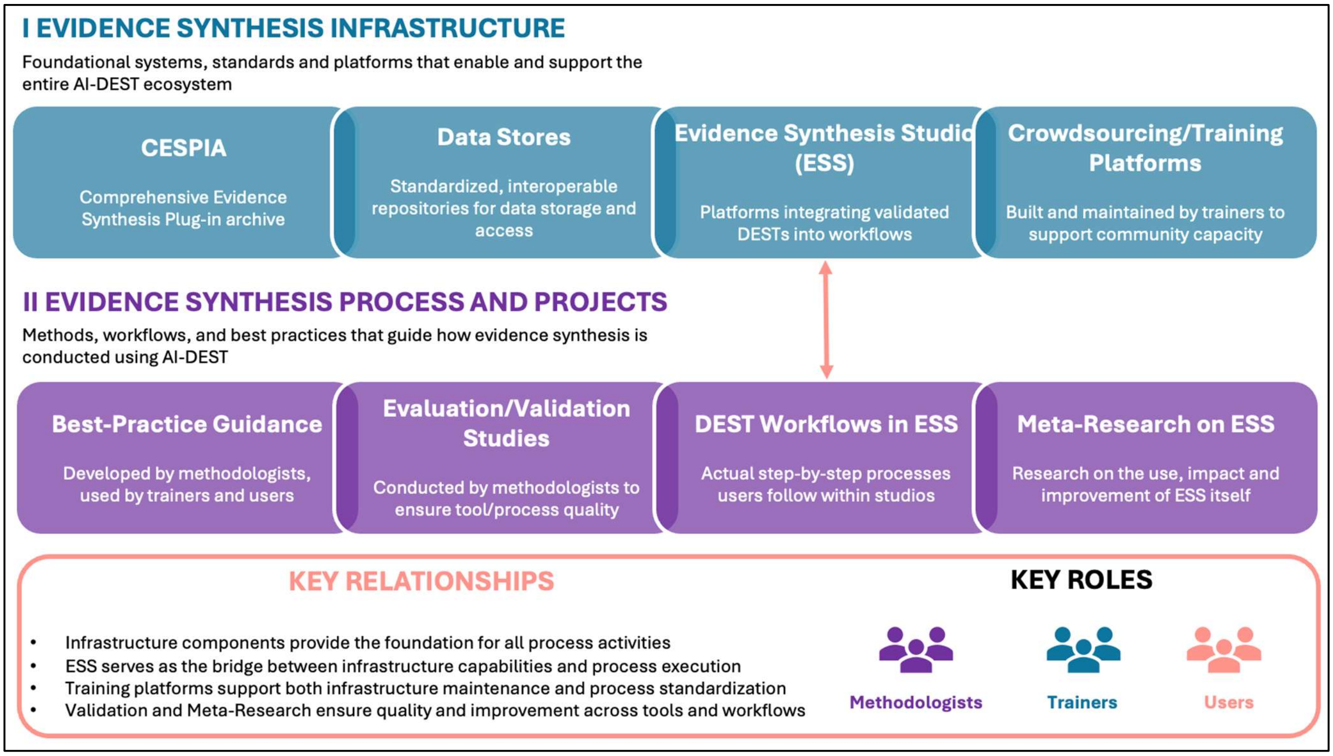
All solutions were rated as high impact. **Evidence Synthesis Studio (ESS)** was rated as the highest impact solution and was unanimously agreed by all the Working Group members.

PROPOSED SOLUTIONS/STRATEGIES

Figure 2 below highlights the relationship between recommendations and depicts an integrated framework for AI-Digital Evidence Synthesis Tools (DESTs), showcasing how different components interact to create a cohesive validation and development ecosystem The interconnected components of the AI-DEST ecosystem work together to enhance evidence synthesis, ensuring efficiency, accuracy, and accessibility.

Evidence Synthesis Infrastructure provides a foundational systems, standards, and platforms that support the entire process. **CESPIA**, the Comprehensive Evidence Synthesis Plug-in archive for AI-DESTs, serves as both a governance framework and a live inventory for validated AI-DEST tools. It plays an **active role** in defining validation standards, maintaining datasets, and ensuring interoperability across evidence synthesis platforms. Rather than simply listing validated tools, CESPIA establishes **rigorous validation protocols**, curates datasets for independent testing, and provides the infrastructure necessary for ongoing assessment of AI-DEST capabilities. This ensures that tools are not only cataloged but also continuously evaluated for their effectiveness in different research contexts. CESPIA's governance function extends beyond validation—it **sets the criteria** for what qualifies as a validated tool, specifying its intended applications and limitations. By maintaining a structured inventory based on independent validation processes, CESPIA ensures that researchers, policymakers, and practitioners can confidently access AI-DESTs that meet established standards. Additionally, it fosters **community-driven validation**, incorporating participatory approaches and peer feedback to refine tools and improve usability across disciplines. Through its **dual role** as a validation authority and inventory manager, CESPIA strengthens transparency, accessibility, and trust in AI-driven evidence synthesis. It ensures that AI-DEST tools are not only available but also **reliable, adaptable, and ethically governed**, supporting interdisciplinary collaboration and informed decision-making. The **ES Data Store** acts as a centralized repository, maintaining structured and synthesized evidence outputs. It ensures interoperability between AI-DESTs, CESPIA, and ESS for seamless data exchange, supporting data standardization, persistent identifiers, and metadata enhancement to make evidence management more efficient. The **Evidence Synthesis Studio (ESS)** serves as a dynamic AI-supported infrastructure, integrating DESTs across various synthesis processes that enables multiple studios to function as adaptable environments for evidence synthesis. ESS allows users to **assemble DESTs in a 'plug and play' manner**, tailoring workflows to their specific review needs. This flexibility ensures that researchers can configure their synthesis processes dynamically, whether working individually in a **studio-like setting**—similar to R Studio—or within a structured team environment where a predefined set of DEST is curated for broader application by downstream users. ESS advances current evidence synthesis platforms by integrating **validated DESTs** into structured workflows, supporting **quantitative, mixed-methods, and qualitative approaches**. It enables policymakers and researchers to **rerun analyses with customized parameters**, refining evidence synthesis outputs to enhance decision-support capabilities. By fostering both **individual customization and collaborative synthesis**, ESS ensures that AI-driven methodologies remain adaptable, scalable, and responsive to diverse research needs. **Crowdsourcing and Training Platforms** contribute to community capacity-building by supporting trainers and facilitating broader adoption of AI-DEST methodologies.

Figure 2. AI-DEST Core components and structure


I EVIDENCE SYNTHESIS INFRASTRUCTURE
Foundational systems, standards and platforms that enable and support the entire AI-DEST ecosystem
CESPIA
Comprehensive Evidence Synthesis Plug-in archive
Data Stores
Standardized, interoperable repositories for data storage and access
Evidence Synthesis Studio (ESS)
Platforms integrating validated DESTs into workflows
Crowdsourcing/Training Platforms
Built and maintained by trainers to support community capacity
II EVIDENCE SYNTHESIS PROCESS AND PROJECTS
Methods, workflows, and best practices that guide how evidence synthesis is conducted using AI-DEST
Best-Practice Guidance
Developed by methodologists, used by trainers and users
Evaluation/Validation Studies
Conducted by methodologists to ensure tool/process quality
DEST Workflows in ESS
Actual step-by-step processes users follow within studios
Meta-Research on ESS
Research on the use, impact and improvement of ESS itself
KEY RELATIONSHIPS
KEY ROLES
Methodologists
Trainers
Users

- Infrastructure components provide the foundation for all process activities
- ESS serves as the bridge between infrastructure capabilities and process execution
- Training platforms support both infrastructure maintenance and process standardization
- Validation and Meta-Research ensure quality and improvement across tools and workflows

The **Evidence Synthesis process and projects** encompasses the methodologies, workflows, and best practices that dictate how AI-DEST is applied in research. **Best-Practice Guidance**, developed by methodologists, provides essential frameworks for trainers and users, ensuring standardization in evidence synthesis. **Evaluation and Validation Studies** assess the reliability and quality of AI-DEST tools and workflows, maintaining high standards across the ecosystem. Within ESS, **DEST Workflows** guide users step by step, enabling a structured approach to conducting synthesis. DEST Development focuses on the iterative refinement and optimization of AI-DEST tools. This includes real-time benchmarking, validation, and sector-specific improvements to enhance AI functionality and ensure adaptability to diverse evidence synthesis needs.

These components work in harmony: Validation feeds into CESPIA to maintain AI-DEST transparency, while DEST Development is refined through validation mechanisms to ensure continuous improvements. CESPIA (R3) connects to ESS (R1), allowing seamless AI-DEST integration into evidence synthesis workflows. ESS interacts with the ES Data Store to ensure that evidence outputs remain structured, accessible, and adaptable.

I. EVIDENCE SYNTHESIS INFRASTRUCTURE

Recommendation 1. Creating an evidence Synthesis Studio [ESS]

Problem Addressed

In stages 1 to 3 of the ESIC process, WG3 identified several gaps and challenges in the current state of AI-DEST including the lack of standardised workflows that facilitate integrated synthesis across different evidence synthesis types and disciplines. Automated evidence packaging that combines quantitative and qualitative synthesis with contextual policy insights is limited, and many AI-DESTs lack features for contextualising, visualising, and translating synthesised results into accessible formats for policymakers. The development of platforms that support layered summaries, multilingual interfaces, and linkage with decision-support systems is limited. AI tools are not optimized for processing grey literature, such as UN evaluation reports, or publications with inconsistent metadata, making their applicability beyond journal articles unreliable. Additionally, concerns persist regarding the reproducibility and consistency of AI-based automation tools, particularly their varying levels of accuracy and efficiency across sectors and contexts. Citation retrieval for grey literature remains a challenge, with NLP tools for natural language queries still under development. Full-text screening tools have lower validation compared to title and abstract screening, limiting accuracy, reproducibility, and contextual understanding in complex documents. Extraction tools exhibit performance variability, making structured data extraction prone to errors and reliant on input quality and domain knowledge. Many existing AI-DESTs have been developed and optimized for randomized controlled trials (RCTs) in health sciences, not representing the broader evidence landscape. Lacks support for qualitative, mixed-methods, case study-based, observational and other types of evidence seen in fields like social sciences, policy research and evaluation, or development studies.

Solution/Strategy

The proposed **Evidence Synthesis Studio (ESS)** serves as a dynamic AI-supported infrastructure, integrating DESTs across various synthesis processes that enables multiple studios to function as adaptable environments for evidence synthesis. ESS allows users to assemble DESTs in a ‘plug and play’ manner, tailoring workflows to their specific review needs. This flexibility ensures that researchers can configure their synthesis processes dynamically, whether working individually in a studio-like setting—**similar to R Studio**—or within a structured team environment where a predefined set of DEST is curated for broader application by downstream users. ESS advances current evidence synthesis platforms by integrating validated DESTs into structured workflows, supporting quantitative, mixed-methods, and qualitative approaches. It enables policymakers and researchers to rerun analyses with customized parameters, refining evidence synthesis outputs to enhance decision-support capabilities. By fostering both individual customization and collaborative synthesis, ESS ensures that AI-driven methodologies remain adaptable, scalable, and responsive to diverse research needs.

ESS is designed for flexibility, supporting quantitative, mixed methods, and qualitative synthesis while ensuring inclusivity across research needs. It enables multiverse approaches, allowing multiple frameworks or AI-DESTs per task, with human-in-the-loop oversight for expert validation. A modular, scalable pipeline ensures efficiency, while a citizen science interface fosters public collaboration. Users can **r**e-run analyses, adjusting input restrictions for methodological precision. ESS incorporates open data standards (JSON, XML, CSV) and API endpoints to enable automated workflows and persistent identifiers for reproducibility. It embeds Explainable AI (XAI) for transparent decision-making, tracks computational usage for sustainability, and integrates **CESPIA**, a validated tool repository.

AI-powered search and screening, bibliometric tracking, and thematic clustering refine evidence synthesis, ensuring usability, adaptability, and continuous optimization through feedback and automation

Justification

The **Evidence Synthesis Studio (ESS)** is envisioned as a modular, scalable, and automated platform that enables researchers to conduct systematic reviews, evidence mapping, and meta-analyses with efficiency and precision. Rather than a singular entity, ESS represents a flexible framework where multiple studios can be configured to meet diverse research needs. Researchers will be able to assemble DESTs in a 'plug and play' manner, tailoring workflows to their specific synthesis requirements. ESS will integrate AI-driven processes with human oversight to enhance reliability, accessibility, and adaptability in knowledge synthesis.

Innovation: Evidence Synthesis Studio (ESS) serves as a platform for integrating validated DESTs into structured workflows, ensuring seamless application in evidence synthesis processes. Some ESS providers will offer a customized and pre-determined user experience, streamlining workflows for specific research needs. Others will provide a more flexible, bare-bones interface, similar to R Studio, allowing users to tailor their approach based on individual preferences and requirements. Additionally, certain ESS providers will adopt a hybrid model, combining structured guidance with customizable features to accommodate a diverse range of research methodologies and user expertise. This adaptability ensures that ESS remains a versatile tool for researchers, supporting both standardized and highly personalized evidence synthesis approaches.

Expected Outcomes and Impact: Accelerated evidence synthesis, reducing review time while maintaining rigor. Increased accessibility supporting multilingual users and adaptive interfaces.

Vision of Success (S.M.A.R.T.): The vision is that in 3–5 years, the Evidence Synthesis Studio (ESS) will be a fully operational and widely adopted platform that integrates customizable AI-assisted synthesis workflows, significantly enhancing the efficiency and accessibility of evidence synthesis.

- *Specific:* Establish a fully functional ESS with customizable AI-assisted synthesis workflow
- *Measurable:* Achieve at least 50% reduction in synthesis time compared to traditional methods.
- *Achievable:* ESS will gain widespread acceptance among researchers and institutions, with a measurable increase in adoption rates across diverse research fields.
- *Relevant:* Align ESS with global systematic review standards, including PROSPERO registration.
- *Time-bound:* The development and scaling of ESS will follow a structured timeline, with full implementation expected within 3–5 years.

Strategic Value

Sequencing and Timing:

- **Phase 1 (0–6 months)**: Development of core ESS infrastructure and AI modules.
- **Phase 2 (6–12 months)**: Beta testing with selected institutions, refining modular architecture.
- **Phase 3 (12–24 months)**: Scaling deployment, integrating multilingual interfaces, expanding CESPIA validation.

Alignment with broader infrastructure: Complying with open science and FAIR data principles (Findable, Accessible, Interoperable, Reusable).

Value Assessment: Efficiency gains by reducing resource-intensive manual research tasks. Scalability, ensuring compatibility across diverse research domains. Sustainability, adopting environmentally conscious computing practices

Equity: Supporting diverse research institutions, particularly in low-resource settings and prioritizing inclusivity in multilingual

Legitimacy: Establishes rigorous validation standards, ensuring AI-DEST tools meet credibility benchmarks for informed decision-making.

Systems Approach: ESS adopts a holistic evidence synthesis model that integrates AI-assisted workflows for efficiency, human oversight for methodological rigor, automated surveillance for continuous evidence updates, and scenario planning tools (dashboard!) for policy forecasting

Recommendation 2. An Inventory of AI DESTs [CESPIA]

Problem Addressed

The integration of AI in DEST presents challenges in validation, interoperability, transparency, and equitable access. Many AI-driven evidence synthesis tools (AI-DESTs) vary in their effectiveness and lack standardized validation protocols, leading to uncertainty among researchers and policymakers regarding their reliability. Without a centralized governance framework, inconsistencies in methodology and trustworthiness limit the adoption and usability of these tools across disciplines. Limited transparency, along with missing metadata and validation processes, also makes it harder to trust and use these tools. Additionally, the lack of standard performance benchmarks means professionals working in synthesis, evidence, and research must spend extra time understanding data science and collaborating with experts to ensure AI applications work properly and deliver reliable results.

Solution/Strategy

CESPIA addresses these challenges by providing a **community-enabled** governance framework that ensures standardized validation, continuous assessment, and interoperability of AI-DEST tools. Rather than acting as a passive registry, CESPIA actively curates datasets, defines validation protocols, and maintains an evolving inventory based on independent testing and peer validation. It bridges AI-DEST tools with the ES Data Store, ensuring structured data management and seamless evidence exchange while incorporating persistent identifiers and metadata enhancement.

Justification

Innovation: CESPIA transforms evidence synthesis by ensuring continuous validation and adaptive governance, distinguishing it from traditional static registries. Its **community-driven validation approach** leverages participatory feedback, enabling interdisciplinary collaboration while refining usability across research domains. By integrating **equity and legitimacy** principles into validation, CESPIA prioritizes ethical AI deployment, ensuring tools meet established trust and transparency benchmarks.

Expected Outcomes and Impact: CESPIA oversees the seamless interaction between datastores, ensuring that multiple repositories communicate effectively to support evidence synthesis. It maintains a comprehensive inventory of validated tools, offering researchers easy access to AI-DESTs, similar to how R packages are downloaded and utilized. Beyond technical validation, CESPIA fosters consensus around tool usage and guidelines, incorporating principles such as environmental impact considerations and ethical AI practices.

- Reliable AI-DEST Tools: Researchers and practitioners gain access to AI-validated evidence synthesis tools, enhancing decision-making across disciplines.
- Interdisciplinary Collaboration: Standardized validation fosters trust across research domains, supporting widespread adoption.
- Ethical AI Governance: CESPIA ensures AI tools operate within established ethical principles, strengthening transparency and reducing biases.
- Improved Data Interoperability: The ES Data Store optimizes data exchange, enabling efficient evidence standardization.

Vision of Success (S.M.A.R.T.):

The **Vision of Success** for CESPIA revolves around fostering collaboration between researchers and developers to refine and validate DESTs, ensuring that only rigorously tested and approved tools are integrated into the evidence synthesis ecosystem. By serving as the foundation for a live inventory of AI-DEST tools, CESPIA will provide a trusted repository that enhances transparency, efficiency, and innovation in research.

- *Specific:* Establishing CESPIA as a global reference for AI-driven evidence synthesis validation.

- *Measurable:* Implementing regular independent audits, ensuring AI-DESTs adhere to validation protocols.
- *Achievable:* Collaborating with AI researchers, policymakers, and institutions to maintain participatory validation.
- *Relevant:* Addressing the gaps in AI governance by reinforcing interoperability, trust, and usability.
- *Time-bound:* Fully integrating CESPIA with AI-DEST tools and the ES Data Store within five years.

Strategic Value

Sequencing and Timing:

- Year 1–2: Developing validation criteria and community governance framework.
- Year 3–4: Curating datasets, refine interoperability mechanisms, and establish independent testing protocols.
- Year 5: Ensuring full-scale implementation with structured AI-DEST inventory and seamless ES Data Store integration.

Alignment: CESPIA aligns AI-driven synthesis with research ethics, ensuring equitable access and global applicability. It will integrate seamlessly with the Evidence Synthesis Studio (ESS) and other AI-DEST platforms, fostering collaboration between researchers, developers, and policymakers.

Value Assessment: CESPIA's value is anchored in trust, transparency, usability, and adaptability. By embedding system-thinking principles, it ensures AI-DEST tools evolve alongside research needs while maintaining rigorous governance. Its structured evidence repository enhances value-driven decision-making, reinforcing CESPIA's role as a foundational framework for AI-driven synthesis.

Equity: Prioritizes inclusive validation processes, incorporating diverse stakeholder feedback.

Legitimacy: Establishing rigorous validation standards backed by independent assessments.

Systems Approach: Ensures continuous validation, interdisciplinary collaboration, ethical AI governance, and seamless integration with the ES Data Store. It fosters adaptive accountability, where AI-DEST tools evolve through peer feedback and independent testing, promoting transparency and trust. By embedding interoperability mechanisms, CESPIA enables structured data management, improving usability and accessibility across discipline.

Recommendation 3. ES Data store/Repository of Evidence

Problem Addressed

The ES Data Store faces several challenges that hinder its efficiency and reliability. Interoperability issues arise due to the lack of standardized data exchange protocols, limited API integration, and compatibility concerns within existing ES software, making seamless data sharing difficult. Reproducibility in Large Language Model (LLM) outputs is another challenge, as their stochastic nature and probabilistic word prediction can lead to inconsistencies—small variations in input, model configuration, or version can result in different outputs, complicating validation and comparison. Additionally, security measures are sporadically and inconsistently adopted, leaving vulnerabilities in data integrity and access control.

Solution/Strategy

The ES Data Store serves as a centralized repository for structured and synthesized evidence outputs, ensuring interoperability between AI-DESTs, CESPIA, and ESS for seamless data exchange. It supports data standardization, persistent identifiers, and metadata enhancement, making evidence management more efficient. To address challenges in interoperability, reproducibility, and security, we recommend the creation of a centralized archive with unique identifiers, systematically storing extracted data and annotations from previous reviews. AI tools can be leveraged to expand the repository, collecting similar information from unselected publications or reports to enhance comprehensiveness. Importantly, this archive will include metadata on annotation provenance, ensuring transparency and traceability. **As part of CESPIA,** the ES Data Store will integrate validated AI-DEST tools, reinforcing trust, accessibility, and ethical governance in evidence synthesis.

Justification

The ESIC planning process is funded by the Wellcome Trust.

Innovation: Future developments will focus on enabling LLM prompts designed for specific tasks or projects to be deployed with confidence in related contexts. This will improve consistency and applicability across different research domains.

Expected Outcomes and Impact: The ES Data Store will enhance data accessibility, interoperability, and security, ensuring that researchers can efficiently retrieve and validate evidence. By improving reproducibility and standardizing security protocols, the repository will become a trusted resource for evidence synthesis.

Vision of Success (S.M.A.R.T.):

- *Specific:* Establish a fully functional ES Data Store with standardized data exchange protocols and robust security measures.
- *Measurable:* Achieve seamless API integration across multiple platforms, reducing interoperability issues by at least 50%.
- *Achievable:* Ensure adoption by a defined number of institutions and researchers, demonstrating widespread usability.
- *Relevant:* Align the repository with global research standards, ensuring compatibility with systematic review frameworks.
- *Time-bound:* Implement core functionalities within 12 months, followed by iterative improvements based on user feedback.

Strategic Value

Sequencing and Timing: The ES Data Store will be developed in phases, ensuring structured implementation and continuous optimization.

- Phase 1 (Year 1–2): Establish data exchange protocols, API integration, and persistent identifiers for reproducibility.
- Phase 2 (Year 3–4): Expand AI-driven metadata enhancement, automate risk-of-bias assessments, and refine interoperability mechanisms.
- Phase 3 (Year 5): Fully integrate with CESPIA, ensuring validated AI-DEST tools and seamless evidence synthesis.

Alignment: The ES Data Store aligns with CESPIA's governance framework, ensuring validated AI-DEST tools meet credibility benchmarks. It supports open data standards, fostering interdisciplinary collaboration and transparent evidence synthesis.

Value Assessment: The ES Data Store enhances evidence synthesis by ensuring structured data management, interoperability, and persistent identifiers for reproducibility. By systematically storing extracted data and annotations, it improves transparency, accessibility, and efficiency, making AI-driven synthesis more reliable. AI tools expand the repository, collecting relevant information from unselected sources, increasing comprehensiveness and usability.

Equity: Ensuring open access and transparency, the repository will be designed to support equitable participation from researchers across diverse backgrounds and institutions.

Legitimacy: As part of CESPIA, the ES Data Store aligns with validated AI-DEST tools, reinforcing credibility and trust in evidence synthesis. It ensures traceability of annotations, allowing researchers to verify provenance and methodological integrity. By integrating standardized validation protocols, it strengthens confidence in AI-assisted decision-making.

Systems Approach: The ES Data Store operates within a dynamic, interconnected framework, ensuring seamless data exchange between AI-DESTs, CESPIA, and ESS. It employs adaptive governance, continuously refining validation standards and interoperability mechanisms. By embedding Explainable AI (XAI) and automated workflows, it supports transparent, scalable, and ethically governed evidence synthesis.

Recommendation 4. Crowdsourcing training platform to support training and adoption

Problem Addressed

There is a need to address the digital divide. Disparities in technology access and digital literacy (concentrated among men, urban residents, and the highly educated) limit public engagement with AI in the Global North, while in the Global South, deep digital divides and a lack of culturally relevant resources further hinder participation. Capacity building is crucial, involving training programs and developing accessible resources to facilitate adoption.

Solution/Strategy

To foster safe and responsible use of AI in evidence synthesis, we propose a capacity building and inclusive training strategy for AI in evidence synthesis. The strategy includes 1) Workshops, webinars, and a structured mentorship program to support hands-on learning and peer exchange, enabling teams to build internal capacity and integrate AI tools into their workflows; 2) Online training modules and practical resources that explain the use and implications of AI in evidence synthesis, tailored for both professionals and the public to promote understanding and responsible engagement; 3) Regular training sessions and strategic promotion efforts to ensure broad participation and sustained impact, with a focus on addressing geographic and equity gaps in digital literacy and access.

Justification

Innovation: This strategy establishes a multi-level capacity-building ecosystem that blends human-centered learning with scalable digital tools to support inclusive and ethical adoption of AI in evidence synthesis. It prioritizes equity through localized, accessible training and promotes responsible use by embedding ethical and practical literacy, ensuring sustainable, real-world readiness for AI-DEST tools.

Expected Outcomes and Impact:

- Increased adoption and responsible use of AI tools among researchers and citizens.
- More equitable evidence ecosystem, with increased participation of underrepresented populations and regions.
- Creation of a self-sustained learning community.

Vision of Success (S.M.A.R.T.):

- *Specific:* Deliver a comprehensive capacity-building program that includes regional workshops, online training modules, and a mentorship network to support the adoption of AI tools in evidence synthesis.
- *Measurable:* Train at least 200 participants in the first 12 months (at least 50% from Global South and 50% from non-health sectors).
- *Achievable:* Different organizations, digital platforms and experts will participate in the co-creation and content delivery.
- *Relevant:* The strategy directly supports AI-DEST uptake while aligning with the need for ethical, effective and inclusive use.
- *Time-bound:* 18 months to completion.

Strategic Value

Sequencing and Timing: Initially, foundational training modules should be developed, focusing on AI-DEST fundamentals and best practices. This should be followed by pilot programs to test usability and effectiveness, incorporating feedback from early adopters. As adoption grows, the platform will expand to include advanced training, certification programs, and integration with research institutions. Continuous updates will ensure alignment with evolving AI methodologies and evidence synthesis needs.

Alignment: The platform will align with global AI and evidence synthesis standards, ensuring compatibility with CESPIA and the Evidence Synthesis Studio (ESS). It will integrate with existing research frameworks, supporting interdisciplinary collaboration and adherence to best-practice guidelines. Partnerships with academic institutions, AI developers, and policy organizations will strengthen its credibility and ensure widespread adoption.

Value Assessment: The platform will enhance researcher proficiency in AI-DEST, reducing barriers to adoption and improving evidence synthesis efficiency. Success will be measured by increased engagement, improved research outputs, and the number of trained professionals utilizing AI-DEST tools effectively. The platform will also contribute to the broader AI ecosystem by fostering innovation and collaboration.

Equity: An equity lens is embedded across design and delivery by ensuring multilingual and accessible tools, involving Global South practitioners in leadership roles, and directing support toward underrepresented populations and institutions.

Legitimacy: The platform will draw on the existing crowdsourcing platforms for evidence synthesis and will be built on validated methodologies.

Systems Approach: This will integrate AI-DEST training with CESPIA validation processes and ESS workflows. The platform will support adaptive learning, iterative improvements, and dynamic updates based on user feedback. By fostering collaboration between researchers, developers, and trainers, it will create a sustainable ecosystem for AI-DEST adoption and evidence synthesis advancement.

II. EVIDENCE SYNTHESIS PROCESS AND PROJECTS

Recommendation 5. A framework for validation of DEST performance

Problem Addressed

Developing a framework for validation of Digital Evidence Synthesis Tools (DESTs) will aim to address the lack of a standardized validation protocol for responsible AI tools in evidence synthesis. Currently, individual organizations assess DESTs based on their own criteria, leading to fragmented evaluations and inconsistent benchmarking. With the rapid evolution of AI models, researchers often rely on personal experience or peer recommendations rather than objective metrics, making validation unreliable. The framework will seek to establish uniform accuracy benchmarking to improve transparency and trust in AI-driven synthesis tools. Additionally, this may tackle the biases inherent in AI development, particularly the dominance of the Global North, which marginalizes perspectives from the Global South in research, language representation, and socioeconomic inclusivity. Large Language Models (LLMs) often exhibit Western bias, performing poorly in non-English languages such as African dialects, further excluding Global South researchers. By ensuring robust validation, standardized performance metrics, and inclusivity, the framework strives to create a more equitable and effective AI-driven evidence synthesis ecosystem.

Solution/Strategy

We recommend the development and continuous refinement of a consensus validation framework for DESTs, incorporating standardized definitions of key performance measures to ensure clarity and effectiveness. This framework should thoughtfully integrate considerations of cultural sensitivities, language diversity, and accessibility to promote inclusivity and equitable application. Furthermore, substantial citizen input must be embedded within the process to define validation measures, ensuring that the framework reflects societal needs and expectations. To enhance transparency and reproducibility, the framework should include clear definitions for ‘data statements’ or ‘model cards,’ which comprehensively outline the key features of a DEST. Collectively, these elements will foster trust, accountability, and reliability in DEST evaluations, ultimately supporting their responsible and beneficial deployment.

Justification

Innovation: The framework for validating DEST performance will integrate standardized performance metrics to ensure consistent and reliable assessments across different implementations. It will aim to emphasize substantial **citizen participation in defining validation measures,** fostering transparency and ensuring that the framework aligns with societal needs. To further enhance accountability and reproducibility, the framework will include ‘data statements’ or ‘model cards,’ which provide clear documentation of key features and operational details of DESTs

Expected Outcomes and Impact:

- Improved trust in DEST implementations and their validation process

- Encourages broader adoption due to clear validation criteria

Vision of Success (S.M.A.R.T.):

- *Specific:* Establishing a widely accepted citizen-driven validation framework, which is implemented through structured participatory approaches (such as collaborative testing initiatives, crowdsourcing, feedback loop for continuous improvement)
- *Measurable:* Tracking adoption rate, tracking participation levels in citizen-driven validation processes, including feedback submissions, community review panels, and collaborative testing initiatives.
- *Achievable:* Leveraging existing community engagement models for crowdsourced validation, ensuring broad participation from researchers, practitioners, policymakers, and the general public
- *Relevant:* The framework ensures DEST outputs align with ethical AI principles while remaining adaptable to multilingual, cross-sectoral, and interdisciplinary needs
- *Time-bound:* The Validation Framework for DEST Performance will be time-bound, ensuring continuous refinement through structured phases of development and implementation. Initially, it will establish standardized performance measures, followed by iterative improvements based on stakeholder feedback and evolving AI capabilities. Regular assessment cycles will ensure adaptability to emerging research needs while maintaining rigorous validation protocols.

Strategic Value

Sequencing and Timing: can be done in parallel to ESS development and will feed in the ESS process of validation. Engaging citizens and experts to generate the 'essential dimensions' for the framework (12 months). These 'essential dimension' could include 'performance targets', which are essentially metric that may involve a) precision and recall of retrieved sources (such Introduce a standardized protocol for testing, covering aspects such as Search & retrieval efficiency, Screening & extraction accuracy, Synthesized output coherence, Bias detection & transparency

- Defining acceptance criteria based on use cases for different sectors and scenarios.
- Defining threshold for each acceptance criteria. Applying/testing the framework 3-6 months: Selecting the tools and pilot testing the tools for the performance targets: This will involve funding for research projects and developing engagement strategies, such as crowdsourcing models, collaborative forums, and expert review panels, to facilitate testing contributions.

Alignment: In alignment with CESPIA, the framework will integrate community-driven validation, ensuring DEST tools meet credibility benchmarks and remain transparent, equitable, and interoperable within the broader AI-driven evidence synthesis ecosystem. By embedding traceability mechanisms and inclusive validation processes, it will reinforce trust, accessibility, and ethical governance in AI-assisted decision-making.

Value Assessment: Enhances AI-DEST accountability and reproducibility. The value is in the trust in the tools meeting the standards set through the process which would encourage adoption of validated tools

Equity: Ensures inclusive participation through the involvement of diverse groups in the validation process including citizen engagement.

Systems Approach: synergies with CESPIA to validate the tools added to the repository and a standardized approach to assessing new tool

Recommendation 6. Follow best practices re governance, guidelines and ethical use of AI-DEST

Problem Addressed

Ethical guidelines often face limited application due to low awareness, causing inconsistent adherence. Existing frameworks fail to bridge digital divides, leading to unequal AI-driven evidence access. Research in DESTs is largely from the Global North, limiting diversity. The absence of global regulations adds uncertainty in standards and accountability. Additionally, GenAI development raises environmental and human rights concerns, including exploitative labor practices. To resolve these gaps, global AI governance should be led by norm-setting organizations like Cochrane and Campbell, promoting ethical, inclusive, transparent, and sustainable AI development without rigid regulations.

Solution/Strategy

Efforts should be consolidated to establish ethical AI guidelines for evidence synthesis, ensuring clear stakeholder roles, data privacy, and compliance. These guidelines must define responsibilities, address biases, and integrate RAISE with other frameworks. Transparency should be reinforced through "data statements" or "model cards," detailing sources, methods, and biases. Accountability mechanisms and global compliance frameworks are essential for responsible AI use. Ethical guidelines should explicitly cover data privacy and bias mitigation. Citizen science platforms should involve the public in screening tasks, with new models recognizing contributions. Community-led councils should oversee AI systems to ensure transparency, fairness, and cultural relevance.

Justification

Innovation: The innovation lies in its adaptive, participatory, and globally harmonized approach to AI governance, going beyond existing frameworks like RAISE and other ethical guidelines on AI-DEST. We propose to integrate real-world stakeholder engagement, ensuring citizen science platforms, community-led councils, and interdisciplinary collaboration actively shaping AI-DEST governance.

Expected Outcomes and Impact: Improved consistency in AI-DEST applications, reducing bias and enhancing reproducibility. Policymakers will gain confidence in AI-driven outputs, facilitating informed decision-making. The broader impact includes increased adoption of AI methodologies across disciplines, fostering innovation and collaboration.

Vision of Success (S.M.A.R.T.):

- *Specific:* Implementing a comprehensive governance framework for AI-DEST, ensuring ethical AI deployment, transparency, and accountability in evidence synthesis.
- *Measurable:* Monitoring compliance rates, stakeholder engagement, and AI-DEST adoption through performance audits, bias assessments, and reproducibility tracking.
- *Achievable:* Interdisciplinary collaboration to refine ethical guidelines, ensuring scalability and adaptability.
- *Relevant:* Addressing bias mitigation, equitable AI access, and responsible AI governance, ensuring AI-DEST tools align with global ethical standards.
- *Time-bound*: Achieving full-scale implementation within five years, with iterative improvements based on stakeholder feedback and evolving AI regulations.

Strategic Value

Sequencing and Timing: To be integrated with other solutions

Alignment: Aligns with CESPIA, AI-DEST validation protocols, and the ES Data Store, ensuring seamless ethical governance.

Value Assessment: Improves trust, transparency, and usability in AI-driven synthesis, ensuring structured governance enhances reproducibility and accessibility.

Equity: Prioritizes inclusive validation processes, incorporating diverse stakeholder feedback to ensure equitable AI deployment.

Legitimacy: Strengthens credibility ensuring AI-DEST tools meet rigorous ethical and methodological standards.

Systems Approach: Operates within a dynamic, interconnected framework, ensuring adaptive governance, interdisciplinary collaboration, and continuous refinement.

Recommendation 7. Meta-research on the ESS

Problem Addressed

There is a need to determine whether certain levels of error have minimal or negligible impact on the conclusions of an ES project and to assess how errors compound across various stages. Without a clear understanding of these factors, researchers may struggle to evaluate the reliability of synthesized evidence, potentially leading to flawed interpretations and policy recommendations.

Solution/Strategy

To enhance the effectiveness of evidence synthesis and its integration into policymaking, resources should be allocated for targeted research, specifically focusing on error tolerance at different stages of the ES pipeline. This initiative will feed into CESPIA and ESS by:

- Establishing benchmarks for acceptable error levels to ensure synthesized evidence remains robust.
- Developing methodological strategies to assess and mitigate the compounding effects of errors across the ES process.
- Exploring barriers and facilitators that influence the adoption of ES outputs in policy environments.
- Investigating presentation formats that optimize the usability of ES findings, increasing their impact on decision-making.

Justification

Innovation: This will incorporate advanced AI-driven validation techniques and dynamic error assessment models, enabling researchers to quantify tolerable error levels at different stages of evidence synthesis.

Expected Outcomes and Impact: The introduction of structured error tolerance assessment will enhance the reliability of ES outputs, reducing misinterpretations. By improving the accessibility and clarity of ES outputs, resulting in higher adoption rates.

Vision of Success (S.M.A.R.T.):

- *Specific:* Developing and implementing AI-assisted validation tools and structured error assessment frameworks to enhance ESS reliability and usability.
- *Measurable:* Development of error tolerance benchmarks across different stages of evidence synthesis pipelines. Creation of standardized error assessment protocols.
- *Achievable:* Utilizing the existing research framework within the proposed CESPIA and ESS, engaging stakeholders for testing and refinement, and leveraging available funding and resources to support these studies.
- *Relevant:* Addresses fundamental concerns in evidence synthesis, reliability and AI-DEST implementation.
- *Time-bound:* We recommend this as an ongoing activity across the solutions. Conducting foundational research on error tolerance benchmark. Implementing pilot studies to assess the usability of ES outputs. Refining methodologies based on stakeholder feedback.

Strategic Value

Sequencing and Timing: Pilot testing, stakeholder engagement, and iterative improvements based on user feedback.

Alignment: Fits within CESPIA and ESS's overarching goals, reinforcing evidence quality and usability. Supports cross-disciplinary collaboration, ensuring robust research integration.

Value Assessment: Strengthens credibility of ES outputs, contributing to long term evolution of CESPIA and ESS. Continuous assessment providing quantitative and qualitative measures to evaluate reliability and effectiveness of ES outputs.

Equity: Prevent biases and ensures equal participation

Legitimacy: Transparent research processes ensure accountability and trustworthiness.

Systems Approach: Integrates findings into CESPIA and ESS frameworks, reinforcing systemic improvements and utilizes feedback loops and adaptive methodologies to sustain effectiveness

APPENDIXES

Appendix 1. Line listing of Stage 3 recommendations

1	Adopting Open Standards: Using widely accepted data standards and formats (e.g., JSON, XML, CSV) to ensure compatibility and ease of integration with other tools. For example, using standardized data formats for model training ensures that data from different sources can be aggregated and used for training without compatibility issues.
2	Developing APIs: Creating robust Application Programming Interfaces (APIs) that allow seamless access and interaction with databases and between AI DEST modules, with APIs which are well-documented and support different programming languages.
3	Ensuring Data Licensing: Applying open data licenses for evidence synthesis products, including at the level of annotations attributed to individual primary sources (e.g., CC BY, MIT License) to facilitate sharing and reuse, while ensuring proper attribution.
4	Implementing Interoperability Frameworks: Using frameworks like FAIR (Findable, Accessible, Interoperable, Reusable) to guide the development and management of the databases. To ensure data is findable, use persistent identifiers and metadata for discoverability; provide clear access instructions and use open protocols like HTTPS; adopt standardized data formats and vocabularies for interoperability; offer detailed metadata and open licenses for reusability; and use FAIR-supporting tools and platforms, such as data repositories with enforced metadata standards and persistent identifier.
5	Efficient Validation Mechanisms: Creating mechanisms that allow decentralized model validation to ensure that AI tools perform effectively on new data or projects, while requiring less human validation. The goal is to streamline the process of validating AI tools, making it less labor-intensive and enhancing the synergy between human expertise and machine efficiency.
6	Error Tolerance and Research: Commissioning research to understand error tolerance and its impact on evidence synthesis conclusions. This involves studying how resilient the conclusions of an evidence synthesis are to error (human or AI). Understanding error tolerance helps determine the extent to which AI inaccuracies can be accommodated without compromising the integrity of the results. For instance, this would allow identifying the maximum allowable error rate in data extraction that still maintains the validity of the synthesis.
7	Comprehensive Validation Metrics: Establishing quantitative and qualitative validation criteria allows AI tools to be assessed for accuracy, reproducibility, and usability. This provides a structured approach to evaluating AI-driven synthesis models and could usefully be curated in CESPIA.
8	Citizens should have a key role in the establishment of these validation criteria, having their input to deciding what aspects of performance are most important, and the level of performance which is required for different evidence synthesis tasks.
9	Performance Monitoring & Benchmarking: By implementing real-time monitoring metrics within living evidence products, AI tools can be continuously evaluated for effectiveness. A benchmarking system ensures AI solutions are compared against established standards and peer tools, maintaining high-performance levels.
10	Multiverse Approach: Instead of relying on single AI methodologies, this approach enables researchers to compare multiple AI models in parallel. By analyzing outliers and variations, this method ensures that biases are identified, and diverse perspectives are considered, improving evidence integrity. It may be that combining different AI DESTs for the same task (ensemble) provides enhanced performance over any single AI DEST.
11	Human-in-the-Loop Systems: While human validation is currently essential for ensuring reliability, these mechanisms collectively aiming to shift towards an AI-led framework where human oversight is significantly reduced, yet evidence integrity remains uncompromised. Additionally, human-and-AI systems should follow best practices regarding usability to ensure human input is streamlined and efficient. Embedding structured human-in-the-loop workflows not only ensures technical reliability but is also critical for building user trust across diverse evidence consumers.
12	Explainable AI (XAI) in Evidence Synthesis: Enabling transparency in AI systems to ensure users can understand the basis on which decisions were made. Identifying which features most impact the Ai's decisions, helping users see what data the model prioritizes and categorically includes/excludes. Explainability is essential to enable decision-makers, program managers, and community stakeholders to interrogate AI-supported outputs and use them confidently in policy and practice contexts. This might include visual tools like decision trees or heatmaps to illustrate how the model processes information and arrives at conclusions. Using models with clear, interpretable rules allows users to trace the logic behind decisions.
13	To fully leverage XAI in evidence synthesis, it is crucial to establish a basis for evaluating the quality of explanations provided by XAI techniques. Agreed-upon standards and guidelines are necessary to determine what would be considered clear, complete, and accurate, ensuring that these explanations are both useful and trustworthy.
14	Resource Reporting: There is a need for transparency about the computational resources and environmental impact of AI tools. Tracking and reporting the computational power used during AI training and operation, helping assess efficiency. Estimate and report the carbon footprint associated with AI processes, raising awareness about environmental sustainability.
15	Developing a comprehensive, live inventory of AI tools for evidence synthesis, categorized by task and sector, including tools from broader repositories to meet non-health sector needs.
16	Ensure regular updates, including user reviews and performance metrics, and establish CRESPIA, a shared repository of validated tools based on the CRAN model, through collaboration with stakeholders.
17	Developing APIs to integrate AI tools into existing and new evidence synthesis information pipelines.
18	Creating a framework to evaluate AI tools for cultural sensitivity and accessibility, particularly for users in the Global South, ensuring tools can connect with multilingual databases and produce high-quality evidence in multiple languages. Include a live list of languages in which the AI models have demonstrated sufficient proficiency, usually published by the developer of the base language model, to restrict deployment only to languages where strong performance can be expected.
19	Develop AI-supported dynamic evidence mapping systems to reduce time-to-synthesis, enhance transparency, and enable continuous updating of evidence bases.
20	We recommend developing a repository of information which has been derived from primary sources, either in the context of previous evidence synthesis projects or through the opportunistic application of AI DESTs to unselected primary sources, so that these might be (re-) used in other evidence synthesis projects.

wellcome

The ESIC planning process is funded by the Wellcome Trust.

21	Develop an open-source pipeline for evidence synthesis, with desktop- or cloud- based implementations (or both), where users can specify their needs (e.g., type of review, granularity, importance of accuracy).
22	The system provides real-time synthesis pipelines, modular and scalable architecture, cumulative reusable evidence bases, multilingual and accessible interfaces, decentralized governance models, automated evidence surveillance, and environmentally sustainable computing practices.
23	Users can select from different approaches, sources, and tools (e.g., systematic review vs. evidence map, PubMed vs. Open Alex, different risk of bias tools, different AI DESTs).
24	AI is integrated at multiple stages: search, deduplication, screening, PICO annotation, risk of bias annotation, data extraction, and synthesis.
25	The system supports both human and AI-driven processes, with control over how these are selected or combined in each ES stage.
26	Develop the ability for research users to explore an evidence synthesis output by running bespoke analyses specific to their needs on the published output.
27	We recommend the development of automated surveillance functions to detect when evidence syntheses should be updated based on new findings.
28	Automated search tools should be extended to query multiple databases and cover diverse sources, including sector-specific, qualitative, and Global South data. This requires the ability to connect with multiple databases through APIs or direct access protocols, supporting integration across academic, sector-specific, qualitative, and decentralized, locally maintained grey literature repositories, particularly those based in LMICs.
29	AI-powered search string generators using NLP can refine user queries, while vector/semantic search techniques might identify deeper contextual relationships between papers, enhancing retrieval.
30	To enhance screening and extraction, high-performing generic classifiers should be developed for various study designs and research questions.
31	AI-DEST should facilitate automated retrieval of full publications (xml, html, pdf or text) from open sources and institutional subscriptions, bridging performance gaps in screening tools across new domains. This would require integration with APIs of repositories including PubMed Central, Un paywall, publisher websites, JSTOR, OECD and institutional libraries for efficient retrieval.
32	Additionally, an AI-enhanced bibliometric system can track citation trends dynamically, allowing for a deeper understanding of evolving evidence landscapes. Automated surveillance systems should be embedded to detect when significant shifts in evidence accumulation may necessitate updating of existing evidence syntheses, ensuring timely and relevant outputs.
33	AI systems should go beyond structured quantitative evidence to support searching, screening, and extraction across mixed-methods and qualitative research, facilitating automated annotation of qualitative studies with theme recognition, coding, and transcript organization.
34	AI-driven text analysis can cluster recurring themes and insights, while allowing researcher oversight for validation and accuracy.
35	Develop approaches which might allow Large Language Model (LLM) prompts designed for a specific task or project to be deployed with confidence in related tasks or projects.
36	For appraisal and reporting, AI-driven risk-of-bias annotation models should be developed to assess studies against frameworks such as PRISMA, ROBIS, and CASP.
37	AI-DESTs should be developed to support qualitative synthesis summaries, making the interpretation of findings more comprehensive.
38	There should be a mechanism for users to feedback unexpected AI DEST behaviors to the tool developer, so that this may feed into future rounds of AI DEST training; and to CESPIA (see above) so that the performance is visible to others in the community who might use the same tool.
39	Interactive dashboard systems to visualize evidence synthesis outputs for policymakers and practitioners, centralizing data and eliminating fragmentation across sources.
40	The community should be able to nominate tasks for which they would like an AI DEST to be developed.
41	Workshops, webinars, and a mentorship program are recommended to build capacity among researchers and practitioners and support teams in adopting AI tools
42	Inclusive guidelines should address algorithmic and dataset biases, while a harmonized framework should integrate RAISE with other guidelines.
43	These guidelines might include "data statements" or "model cards" for transparency and reproducibility, detailing data sources, gathering methods, model specifics, and potential biases.
44	Establish accountability mechanisms: Define roles and responsibilities for all stakeholders to uphold responsible use.
45	Collaboration with regulatory bodies to establish compliance frameworks that are globally applicable and adaptable.
46	Ethical guidelines must address data privacy and bias, and collaboration with regulatory bodies is essential to establish compliance frameworks that are globally applicable and adaptable.
47	The future of PROSPERO should be secured, with an API endpoint commissioned to allow evidence synthesis pipelines to programmatically verify unique protocol existence and dates.
48	To enhance citizen engagement in AI-DEST, training and resources should be provided, including online modules on AI tools and their applications in evidence synthesis.
49	Regular training sessions could help users understand and use tools effectively and ethically, with strategic promotion to ensure participation, especially in the Global South.
50	Citizen science platforms should be developed to involve the public in tasks including screening and data extraction, and new contribution models should recognize and reward citizen involvement. Community-led citizen governance councils should be established to oversee AI systems, ensuring transparency, fairness, and cultural relevance throughout AI-driven evidence synthesis processes.

Appendix 2. Link between stage 3 and stage 4a reports

	Stage 3 report solutions	R1	R2	R3	R4	R5	R6	R7	Impact Median (IQR)	Effort Median (IQR)
1	Adopting Open Standards	🟪							8,50 (1)	4,50 (1,75)
2	Developing APIs			🟪					8,00 (1,75)	5,00 (1,75)
3	Ensuring Data Licensing								8,00 (1,75)	4,00 (2,75)
4	Implementing Interoperability Frameworks								8,00 (1)	4,00 (2)
5	Efficient Validation Mechanisms	🟪	🟪			🟩			8,00 (1)	5,00 (2,75)
6	Error Tolerance and Research							🟩	7,50 (3)	5,50 (2,50)
7	Comprehensive Validation Metrics		🟪			🟩			7,50 (2,50)	5,50 (3)
8	Citizen's role in validation criteria					🟦			7,50 (1,75)	4,00 (1)
9	Performance Monitoring & Benchmarking		🟦						8,00 (1)	4,00 (2,75)
10	Multiverse Approach	🟦							8,00 (1,75)	4,50 (2,50)
11	Human-in-the-Loop Systems	🟦							8,50 (1,75)	6,00 (3,75)
12	Explainable AI (XAI) in Evidence Synthesis	🟪							8,00 (2)	4,00 (1,75)
13	Agreed-upon standards and guidelines for XAI	🟪							8,00 (1)	4,50 (2,50)
14	Resource Reporting	🟦							7,00 (2,75)	6,00 (2)
15	Live Inventory of AI Tools	🟪	🟩						8,50 (2)	7,00 (3)
16	CESPIA repository modeled after CRAN	🟪	🟩						8,00 (1,75)	5,00 (2,75)
17	Integrating AI into Pipelines	🟩							8,50 (1,75)	4,50 (3)
18	Framework for Cultural Sensitivity & Accessibility					🟦			8,00 (1)	6,50 (3)
19	AI-supported Dynamic Evidence Mapping	🟩							9,00 (2)	6,00 (3,50)
20	Repository of Derived Information			🟩					9,00 (1,75)	5,50 (1,75)
21	Open-source Evidence Synthesis Pipeline	🟩							9,00 (1)	5,50 (2,75)
22	Real-time Synthesis Pipelines	🟩							9,00 (2)	3,00 (3)
23	Users' approaches, sources, and tools selection.	🟦							9,00 (1,75)	5,50 (4,50)
24	AI integration in ES steps	🟩							9,00 (1)	5,00 (2,50)
25	System supports both human and AI-driven processes	🟩							9,00 (2)	7,00 (3)
26	Exploring Outputs with Bespoke Analyses								8,00 (1)	6,00 (2)
27	Automated Surveillance for Updates	🟦							9,00 (1,75)	6,00 (3)
28	Automated Multi-source Search Tools	🟦							9,00 (0,75)	5,50 (2,75)
29	AI-powered Search String Generators	🟦							7,50 (2,75)	5,50 (2,75)
30	Generic Classifiers for Screening	🟦							8,50 (1)	5,00 (3,50)
31	AI for Full Publication Retrieval	🟦							8,50 (2,50)	4,50 (3,75)
32	AI-enhanced Bibliometric System	🟦							8,50 (3,50)	6,00 (3,75)
33	Support for Qualitative Research	🟦							9,00 (2)	5,50 (2)
34	Recurring themes and insights clustering	🟦							7,50 (1,75)	5,00 (2)
35	Reusable Prompts for LLMs			🟦					9,00 (1)	6,00 (3)
36	AI-driven Risk-of-Bias Annotation	🟦							8,00 (2)	7,00 (2)
37	AI Support for Qualitative Synthesis Summaries	🟦							8,00 (1,75)	7,00 (3)
38	Feedback on AI Behavior	🟦							8,00 (2)	7,50 (1,75)
39	Interactive Dashboards	🟦							9,00 (2)	6,50 (2,75)
40	Community Task Nominations								8,00 (2)	7,50 (3)
41	Workshops and Mentorship Programs				🟩				9,00 (2)	7,50 (3,50)
42	Inclusive Guidelines						🟩		8,00 (1,75)	7,50 (3)
43	Include "data statements" or "model cards in guides					🟦	🟦		8,00 (1,75)	7,50 (2)
44	Accountability Mechanisms						🟦		7,50 (2)	6,50 (2)
45	Collaboration with Regulatory Bodies						🟩		9,00 (2)	6,00 (3)
46	Ethical Guidelines for Privacy and Bias						🟩		8,00 (1,25)	7,50 (2)
47	Securing Future of PROSPERO								8,00 (1,25)	6,50 (3)
48	Citizen Training Resources				🟩				8,00 (2)	7,00 (2)
49	Regular training for users				🟩				8,00 (2)	7,00 (2)
50	Citizen Science Platforms						🟦		8,00 (2)	7,00 (4)

🟩	Core solution
🟪	Bases
🟦	Features

The ESIC planning process is funded by the Wellcome Trust.

Appendix 3. Stage 4a Roadmap Integration Table

Problem[i]	Solution			Expected outcome[ii]	Success/ destination[iii]	Synergies[iv]
	Solution name and short description	Type of solution:[v]	Why is the solution innovative?[vi]			
Lack of standardized workflow that facilitates integrated synthesis types, stages and disciplines.	**Evidence Synthesis Studio:** modular, scalable, and automated evidence synthesis platform, enabling researchers to conduct systematic reviews, evidence mapping, and meta-analyses efficiently. It will integrate AI-driven processes with human oversight to enhance reliability and accessibility in knowledge synthesis	(a) evidence synthesis infrastructure (tools or tech / platforms, ongoing products)	Allows users to integrate AI at multiple review stages. **Interest holders impacted**: 1. Policymakers 2. Researchers and Evidence intermediaries 3. Public and Citizens 4. Professionals	- Evidence synthesis time reduction. - Increased accessibility.	A fully functional, customizable AI-assisted Evidence Synthesis Studio (ESS) that reduces synthesis time by at least 50%, achieves widespread adoption across research fields, aligns with global standards, and is fully implemented within 3–5 years.	**Dependencies**: PG1 Core infrastructure quadrant WG2: Depends on adopting open data standards **Complementarity**: WG4 Methods and tools for translating findings for LES to local context; Modular agile toolkit; Establish a global panel of citizen partners, with regional /sub-regional representation; Support the integration of grey literature into evidence synthesis; Pilot "Evidence Response Teams" trained in agile methods and embedded in key institutions (ministries, NGOs) to deliver syntheses within days/weeks of a request). WG5: ESIC Knowledge hub and Knowledge Translation AI (KTai)
No existing database of DESTs making it difficult for researchers to access and evaluate AI tools efficiently.	**An inventory of AI DESTs (CESPIA):** living summary of known AI DESTs, housed within the Comprehensive Evidence Synthesis Plug-In Archive (CESPIA), where users can download AI DEST modules for integration into their Evidence Synthesis Studio pipelines. This system will include real-time benchmarking of performance, reporting of computational and environmental costs, and validation results based on standardized criteria.	(a) evidence synthesis infrastructure (tools or tech / platforms, ongoing products)	Provides a benchmark for assessing effectiveness and accuracy of DESTs, ensuring high quality DESTs integrated into workflows. CESPIA serves as the foundation for a live inventory of AI-DEST tools. **Interest holders impacted**: 1. Researchers and Evidence intermediaries 2. Public and Citizens 3. Professionals 4. Product makers and developers 5. Funders	- CESPIA will increase the number of validated and community-endorsed AI-DEST tools, while enhancing transparency, ethical alignment, and interoperability across evidence synthesis platforms.	CESPIA will be a trusted platform for validating and sharing AI-DEST tools, enhancing transparency and efficiency in evidence synthesis, with measurable growth in adoption and impact within 24 months.	**Dependencies**: PG1 Core infrastructure quadrant WG3: Recommendation 5. A framework for validation of DEST performance and Recommendation 7. Meta-research in ESS **Complementarity**: WG3: Recommendation 1. Evidence Synthesis Studio
Lack of standardized data exchange protocols, limited API integration, and compatibility issues make data sharing difficult.	**ES Data store/ Repository of evidence:** A centralized archive with unique identifiers to systematically store annotations from synthesis. Includes metadata on the provenance of annotations.	(a) Evidence synthesis infrastructure: as this will establish a structured system for assessing and validating DEST performance	Improves consistency and applicability across different research domains. **Interest holders impacted**: 1. Researchers and Evidence intermediaries 2. Product makers and developers 3. Funders	- Increased data accessibility, interoperability, and security.	The ES Data Store will be a secure, standardized repository enabling seamless API integration and reducing interoperability issues by 50%. It will be adopted by key institutions, align with global research standards, and have core features operational within 12 months, with ongoing improvements driven by user feedback.	**Dependencies**: PG1: Engine rooms WG2 4.7 Ensuring quality and Monitoring & Evaluation of data sharing systems **Complementarity**: WG 2 4.1 Living repository of data (includes Open API for data integration, AI-enabled tagging and content structuring)
Disparities in technology access and digital literacy that limit the adoption of AI-	**Crowdsourcing training platform to support training and adoption**: A platform for capacity building and	(a) evidence synthesis infrastructure (tools	Multi-level capacity-building ecosystem that blends human-centered learning with scalable digital tools to support	- Increased adoption and responsible use of AI tools	Deliver a comprehensive, 18-month capacity-building program with workshops, online modules, and	**Dependencies**: PG1: Participatory platforms

DESTs in evidence synthesis processes.	inclusive training that includes online modules, workshops, webinars and a structured mentorship program.	or tech / platforms, ongoing products) (b) evidence synthesis process (methods; training; learning; sharing; convening)	inclusive and ethical adoption of AI in evidence synthesis. **Interest holders impacted**: 1. Policymakers 2. Researchers and Evidence intermediaries 3. Public and Citizens 4. Professionals	among researchers and citizens. - More equitable evidence ecosystem, with increased participation of underrepresented populations and regions. - Creation of a self-sustained learning community.	mentorship to support ethical and inclusive AI-DEST adoption. Train at least 200 participants (50% from the Global South and 50% from non-health sectors) with content co-created by diverse organizations and experts.	**Complementarity**: WG1 strategy: Implementation support to intermediaries (includes mentoring, training, and technical assistance) WG4: Building an academy for evidence synthesis; Incentivize and enhance cross-sectoral learning and collaboration WG5: Mentorship and Train the Trainer program; Continuous Professional Development Modules; and Regional collaborating centers with country nodes
Lack of standardized validation protocol for responsible AI tools in evidence synthesis.	**A framework for validation of DEST performance:** Development and continuous refinement of a consensus validation framework for DESTs, integrating EDI considerations and citizen input.	b) evidence synthesis process (methods; training; learning; sharing; convening) (c) projects (time-limited activities generating outputs)	The validation framework introduces standardized performance metrics and emphasizes citizen participation in defining them, promoting transparency and societal relevance. It also incorporates tools like data statements and model cards to enhance accountability and reproducibility. **Interest holders impacted**: 1. Researchers and Evidence intermediaries 2. Public and Citizens 3. Product makers and developers 4. Funders	- Improved trust in DEST implementations and their validation process - Encourages broader adoption due to clear validation criteria	The goal is to establish a widely accepted, citizen-driven validation framework for DEST performance that is participatory, measurable through adoption and engagement rates, achievable via existing community models, relevant to ethical and cross-sectoral needs, and time-bound with structured phases for continuous refinement and impact assessment.	**Dependencies:** PG1 Core infrastructure quadrant **Complementarity:** WG3: Recommendation 2. CESPIA
Limited implementation of Safe and Responsible use of AI guidelines. Many existing frameworks exist but failure to address digital divides. Absence of global regulatory frameworks, environmental and human right concerns about GenAI.	**Follow best practices re governance, guidelines and ethical use of AI-DEST:** consolidating efforts to integrate RAISE with other ethical frameworks ensuring a structured approach to stakeholders' roles, data privacy and regulatory compliance.	b) evidence synthesis process (methods; training; learning; sharing; convening) (c) projects (time-limited activities generating outputs)	Adaptive, participatory and globally harmonized approach to AI governance. **Interest holders impacted**: 1. Policymakers 2. Researchers and Evidence intermediaries 3. Public and Citizens 4. Professionals 5. Product makers and developers 6. Funders	- Improved consistency in AI-DEST applications, reducing bias and enhancing reproducibility. - Increased adoption of AI methodologies across disciplines.	The goal is to implement a comprehensive, ethically grounded governance framework for AI-DEST that ensures transparency, accountability, and bias mitigation, is measurable through audits and engagement metrics, achievable through interdisciplinary collaboration, and fully implemented within five years with ongoing refinement.	**Dependencies:** PG1: Core infrastructure **Complementarity:** WG3: Recommendation 1 Creating an Evidence Synthesis Studio [ESS] and Recommendation 5. A framework for validation of DEST performance
There is a need to understand error tolerance of AI-DESTs outputs and their impact on evidence synthesis results.	**Meta-research on the ESS:** Funding for research on error tolerance at different stages of ES pipeline, barriers and facilitators of ES uptake into policy, optimization of dissemination strategies for evidence informed policy.	(c) projects (time-limited activities generating outputs)	Incorporates advanced AI-driven validation techniques and dynamic error assessment models, enabling researchers to quantify tolerable error levels at different stages of evidence synthesis. **Interest holders impacted**: 1. Policymakers	- Higher adoption rates of AI-enabled synthesis methods.	The aim is to develop and implement AI-assisted validation tools and error assessment frameworks that enhance ESS reliability, guided by standardized benchmarks and protocols, achieved through stakeholder collaboration and existing infrastructures, and pursued as an ongoing, iterative process to	**Dependencies:** PG1: Core infrastructure WG3: Recommendations 2, and 5 may be adjusted/reshaped based on the results of this recommendation. **Complementarity**: WG3: Recommendation 1 Creating an Evidence Synthesis Studio [ESS]

			2. Researchers and Evidence intermediaries 3. Public and Citizens 4. Professionals 5. Product makers and developers 6. Funders		address core concerns in evidence synthesis.	

[i] What is the problem the solution is intended to address or resolve?

[ii] What are the expected changes or benefits to result from implementing the solution (How does it contribute to cheaper, faster, better synthesis or the SHOW ME principles? Mention the specific type of interest-holders that will be impacted. Link back to the problem statement.

[iii] What would success look like if this solution was implemented? Please be as S.M.A.R.T. as possible.

[iv] What are the connections to solutions proposed by other WGs. Dependency: (development/implementation of the solution is dependent upon another WG solutions; Complementarity: (the solution is helpful or reinforcing in operation. It will have a symbiotic relationship with other solutions).

[v] Type of solution (a) evidence synthesis infrastructure (tools or tech / platforms, ongoing products); (b) evidence synthesis process (methods; training; learning; sharing; convening); (c) projects (time-limited activities generating outputs)

[vi] Why is this solution novel, advantageous, or scalable? (e.g. it’s a novel approach that deviates from conventional methods; it presents clear advantages over existing options; it presents a scalable approach that can extend benefits beyond initial scope?